SUPER: Study of User Priorities for e-Infrastructure for e-Research

Jennifer M. Schopf

Steven Newhouse

Andrew Richards

Malcolm Atkinson

open middleware
infrastructure
institute uk

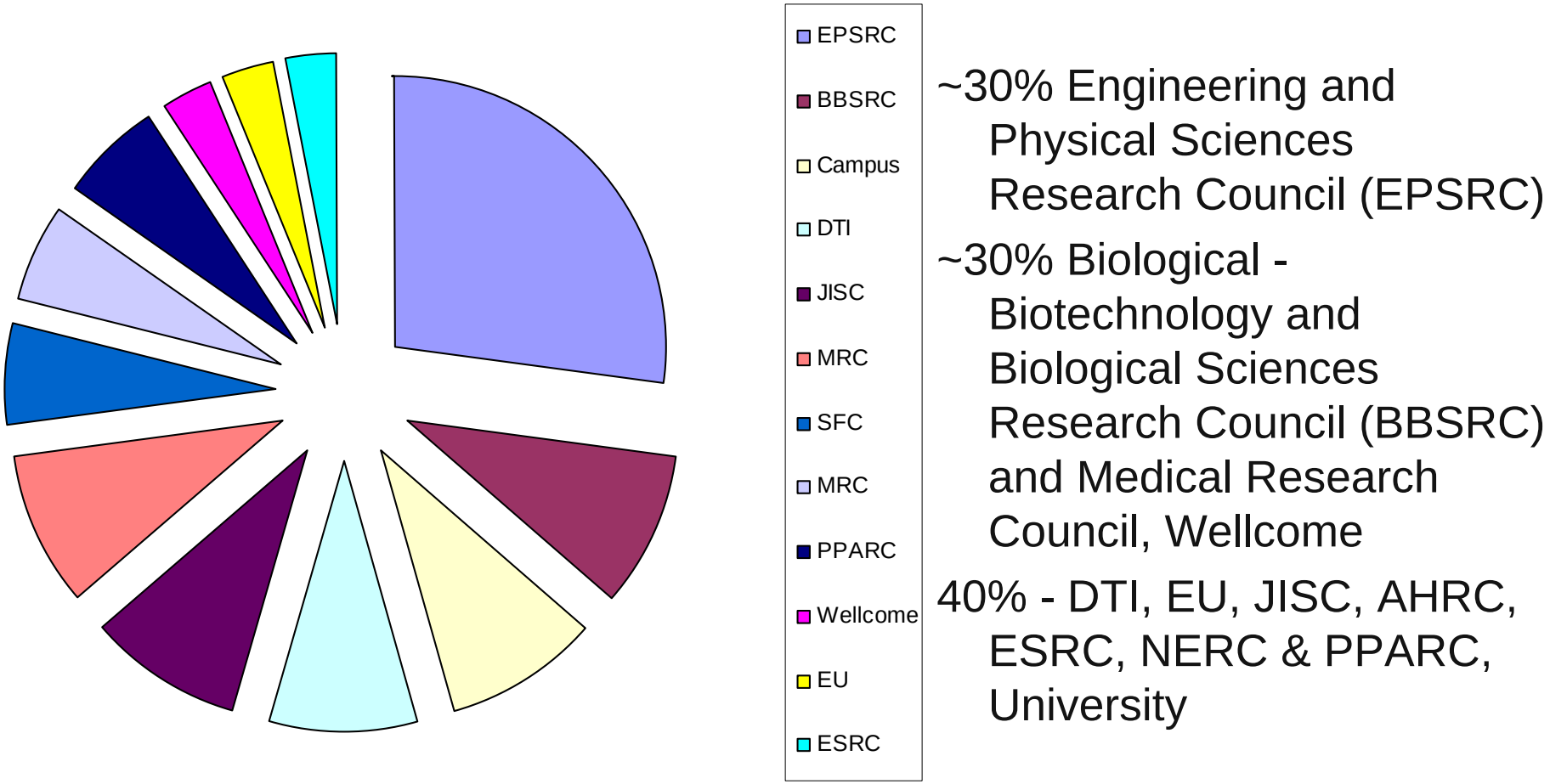
SUPER: Study of User Priorities for e-Infrastructure for e-Research

- Identify issues that are:
  - Short-term (6-18 months):
    - *Actions within existing funding streams*
  - Longer-term (3-5 years):
    - *Actions that need new/renwed funding streams*
- Inform roadmaps for collaborative research:
  - Organisations: OMII-UK, NGS, DCC, …
  - Funders: RC UK, JISC, JCSR, …
- Not the place to identify solutions

Methodology & Coverage

- Face to face interviews:
  - 7 sites - Newcastle, Glasgow, Edinburgh, Oxford, Cambridge, UCL, Reading
  - 45 people from over 30 projects
  - Mostly practitioners
  - Unstructured interviews
- Online survey December 2006 to March 2007
  - ~25 responses
  - 1/3 PIs, 1/3 management, 1/3 “workers”
- One day workshop at NeSC
  - ~30 attendees, mostly funders and PIs

Interview Projects by Funders



~30% Engineering and Physical Sciences Research Council (EPSRC)

~30% Biological - Biotechnology and Biological Sciences Research Council (BBSRC) and Medical Research Council, Wellcome

40% - DTI, EU, JISC, AHRC, ESRC, NERC & PPARC, University

Major Common Topics

- Distributed file management and policy
- Tools to support dynamic Virtual Organisations
- Long-term project support:
  - Teams, services, and training/outreach
- User-Oriented Operational Issues:
  - Authentication, software licensing, and reliable consistent environments
- User Interaction with e-infrastructure services

Sharing Large-Scale Data

- By far the largest concern of the users we spoke with
- How to share data with colleagues
  - Within their project or their wider community
  - Software, results, or other data
- Long-term storage and curation
  - Annotate files with metadata about the contents and provenance
  - Support search and reanalysis at a later date

Metadata is Key to Sharing

- Additional tools are needed to autogenerate metadata
  - How, where, and by what means those data were generated, i.e. their provenance
- Navigate and analyze such data
- If users are responsible for the annotation of their data, the task is generally left undone
  - Often variable quality of self-done annotations, and much variance from practitioner to practitioner
- Automated collection of basic metadata is seen as a large step forward from current practice
  - For some domains specialists may be required

Lack of Metadata Standards

- Well recognized that standards are needed for interop, and acceptance from communities to use standards where they existed
- Standards exist for basic properties
  - Timestamps, basic data collection, some very general metadata sets available
  - Sometimes several standards
  - General acceptance of these
- Lower-level and domain-specific metadata standards are lacking
  - Many communities having to create their own
  - Competing standards are likely – although generally this is acknowledged and they would like to avoid

Longer Term Storage

- Shift towards much longer-term storage of data
  - Up to 40 years for some groups
  - Some for pragmatic experimental use
  - Some at the behest of the funding agencies
- Need for policy discussions in most groups
  - Need to consider both user roles and temporal constraints
- Ned better understood access control mechanisms

Easier Access to Own Data

- Groups now have to manage the file output from computations across multiple locations
  - National resources as well as campus and desktop resources
- Would like to access their local files seamlessly when running an application remotely
  - Edit locally the input files that form the basis of the simulation
  - Output files residing on a remote resource need to be accessible for future processing and analysis on the local desktop
- Requirements for the registration and discovery of files held in different locations

Major Common Topics

- Distributed file management and policy
- Tools to support dynamic Virtual Organisations
- Long-term project support:
  - Teams, services, and training/outreach
- User-Oriented Operational Issues:
  - Authentication, software licensing, and reliable consistent environments
- User Interaction with e-infrastructure services

Virtual Organizations (VOs)

- Currently no clear definition of a VO
  - Makes it harder to understand problems, design tools
- Undetermined factors include
  - How dynamic the VO is
  - If it involves changes in membership of people or in resources
  - Top down vs bottom up
- Different tool requirements

Current VO Tools

- Current VO tools address relatively static VOs and driven from the centre, e.g. VOMS
  - Often solve problems that aren't what the user is interested in
  - Can be difficult to understand or use in production settings
  - Needs to evolve to more than just a management of roles
- Most are built for system administrators
  - Interfaces suited to technically experiences
- Needs basic end-user tools as well
  - What resources they have access to
  - How many cycles they have left as part of a collaboration
  - Easier tools for collaboration and sharing
  - Without in-depth information about certificate setup procedures, for example

Major Common Topics

- Distributed file management and policy
- Tools to support dynamic Virtual Organisations
- Long-term project support:
  - Teams, services, training and consultancy
- User-Oriented Operational Issues:
  - Authentication, software licensing, and reliable consistent environments
- User Interaction with e-infrastructure services

Teams

- Effectively organised team is critical to a project's success
  - Managing distributed teams as one is very hard
  - Manage different cultures, organisations & incentives
- Skills and roles needed by such a team will likely vary over time
  - Lack of available specialists in the necessary CI fields
    - *e.g., Web services, HPC programming, application performance tuning and Java programming)*
  - Hard to find personnel that cross application-technology border

Project Support: Services

- Bridging issues between low-level infrastructure supplied by NGS/campus and community-specific software
  - Community : MIMAS, EDINA, myGrid, …
- No single software provider provides the end-to-end solution needed by every group
- Integration is a key role successful projects consider
- Infrastructure providers are unlikely to be of direct use to particular domains or projects
  - Additional higher-level, domain-focused services needed
- Need for sw consultants to aid in decision making
  - Still need broad outreach and evangelizing about what exists – not what MAY exist in the future

Training/Outreach

- Different training for different stakeholders
  - End-users who will using the e-infrastructure services through the tools and applications
  - Developers, both of generic and domain specific services, who will be using the deployed e-infrastructure services to build other services and tools
  - Deployers with responsibility of managing the required e-infrastructure
- Training materials are needed in many forms
  - Formal instructor-led courses, standalone self-help material, worked examples, reference systems, etc.
  - Many infrastructure projects noted the need for training and had funds – but were ignorant of existing training materials and as a consequence there was considerable duplication of activity

Major Common Topics

- Distributed file management and policy
- Tools to support dynamic Virtual Organisations
- Long-term project support:
  - Tools, services, training and consultancy
- User Oriented Operational Issues:
  - Authentication, software licensing, and reliable consistent environments
- User Interaction with e-infrastructure services

Authentication

- Certificates adopted by service providers
- Very difficult for many end-user communities
  - Deployment of many wrappers around certificates
- This has been a known problem since we started using certificates, and it's still not resolved
  - Partly caused by disconnect between the needs for security vs the needs for usability

Licensing

- Growing use of third party commercial applications, e.g. Matlab
  - In some cases no open source alternative is available
  - Community dependence on certified, licensed software
- No good solution for managing the shifting of licenses on machines across a given site
  - Let alone within a full VO

Reliability

- Mentioned by almost every group we spoke with
  - Even professionally hosted Web services lacked stability
- Many groups simply do not expect that the services will be up; they just work with whatever they find responsive at run-time
- SW issues
  - error or recovery status may not be propagated back to the client application invoking the software
- Service providers change deployments which affect end users unpredictable
- Monitoring software may give conflicting results

Major Common Topics

- Distributed file management and policy
- Tools to support dynamic Virtual Organisations
- Long-term project support:
  - Tools, services, training and consultancy
- User Oriented Operational Issues: authentication, software licensing, and reliable consistent environments
- User Interaction with e-infrastructure services

User Interaction with e-infrastructure Services

- Interactions MUST match the user
  - Technical Expertise
  - Normal Environment
- Command line shells
  - Traditional ‘expert’ interface to systems
- Scripting Environment
  - From within basic shells: Bash, Tcsh, …
  - Application Environments: Perl, Matlab, Python, …
- Workflows
- Portals

Conclusions

- Software
  - Much has been prototyped… but needs hardening and support
  - Reliability a prime concern
  - Ongoing work needs communication between groups
- Policy
  - Need ‘better joined up’ -ness through best practice
  - Data, VOs, Environments, …
- Community Support
  - Still need to tell projects what’s feasible
  - Self-help training materials and hands-on tutorials delivered by trainers for common tools

Previous "Roadtrip"

- Requirements gathering for OMII and Globus
  - July 2005, 25 groups
- Results
  - Still struggling with basic functionality
  - Higher-level services that many middleware groups are concentrating on aren't of interest (yet)
  - Installs are hard
  - Reliability is poor
  - Need training, ongoing discussions between tool builders and end users

How have things changed?

- Still struggling with basic functionality
- Higher-level services that many middleware groups are concentrating on aren't of interest (yet)
- Installs are hard
- Reliability is poor
- Need training, ongoing discussions between tool builders and end users

- Basic job execution not a concern
  - Data and archiving foremost
- Some higher-level tools in use
  - Still need more basic development
- Installs much improved
- Reliability still a problem
- Additional training and outreach still needed

Acknowledgements

- Travel Funding
  - E-Science Core Programme & JISC
- Day release
  - OMII-UK, NGS, JISC, Globus (NSF, DOE)
- Contributors
  - Interviewees for their honesty & flexibility
  - Comments from the community

Further Information

- Jennifer Schopf
  - jms@mcs.anl.gov
  - http://www.mcs.anl.gov/~jms
- Study of User Priorities for e-Infrastructure for e-Research (SUPER)
  - S. Newhouse, J. M Schopf, A. Richards and M.P. Atkinson
  - Tech Report UKeS-2007-01, Apr 07
  - http://www.nesc.ac.uk/technical_papers/UKeS-2007-01.pdf
  - Summary version in the UK eScience Conference, 2007
- Grid User Requirements – 2004: A Perspective From the Trenches
  - Jennifer M. Schopf and Steven J. Newhouse
  - to appear, special issue Cluster Computing Journal, 2007
  - http://www-unix.mcs.anl.gov/~schopf/Pubs/ukuser1-clusterj-07.pdf